Original Article

Two-Step Likelihood Ratio Test for Item-Level Model Comparison in Cognitive Diagnosis Models

Miguel A. Sorrel,[1] Jimmy de la Torre,[2] Francisco J. Abad,[1] and Julio Olea[1]

[1]Department of Social Psychology and Methodology, Universidad Autónoma de Madrid, Spain
[2]Faculty of Education, The University of Hong Kong, Hong Kong

Abstract: There has been an increase of interest in psychometric models referred to as cognitive diagnosis models (CDMs). A critical concern is in selecting the most appropriate model at the item level. Several tests for model comparison have been employed, which include the likelihood ratio (LR) and the Wald (W) tests. Although the LR test is relatively more robust than the W test, the current implementation of the LR test is very time consuming, given that it requires calibrating many different models and comparing them to the general model. In this article, we introduce the two-step LR test (2LR), an approximation to the LR test based on a two-step estimation procedure under the *generalized deterministic inputs, noisy, "and" gate* (G-DINA) model framework, the two-step LR test (2LR). The 2LR test is shown to have similar performance as the LR test. This approximation only requires calibration of the more general model, so that this statistic may be easily applied in empirical research.

Keywords: cognitive diagnosis models, model comparison, item fit, Type I error, power

Cognitive diagnosis models (CDMs) have received increasing attention within the field of educational and psychological measurement. These models are useful tools to provide diagnostic information about examinees' cognitive profiles in domains such as education (e.g., Lee, Park, & Taylan, 2011), measurement of psychological disorders (e.g., de la Torre, van der Ark, & Rossi, 2015), and competency modeling (e.g., Sorrel, et al., 2016). Selection of an appropriate CDM is based in part on model-data fit. Model-data fit can be assessed at the test level (e.g., Chen, de la Torre, & Zhang, 2013; Liu, Tian, & Xin, 2016). If the model, particularly if it has a general formulation, fits the data, then it may be useful to study hypothesis about differences in response processes across items. Relative fit indices can be used to evaluate the discrepancy among different statistical models. According to a recent evaluation on the performance of various goodness-of-fit statistics for relative fit evaluation at the item level, the likelihood ratio (LR) test is more robust than other statistics (Sorrel, Abad, Olea, Barrada, & de la Torre, 2017).

The current implementation of the LR test is very time consuming, given that it requires to calibrate many different models and compare them to the general model. For this reason, the Wald (W) test (de la Torre & Lee, 2013) is generally preferred. In light of this, the primary purpose of this study is to investigate the performance of an approximation to the LR test, the two-step LR (2LR) test, which only requires to estimate the more general model once. Reduced model item parameters are estimated at the item level (i.e., one item at a time) rather than at the test level (i.e., all the items in the test simultaneously) following an alternative, heuristic estimation procedure originally introduced by de la Torre and Chen (2011) and further explored here. This procedure is based on the the *generalized deterministic inputs, noisy, "and" gate* (G-DINA; de la Torre, 2011) model framework. The rest of the article is structured as follows. Next section provides background information about CDMs and model comparison. The design of the simulation study is described thereafter. Subsequently, some results are presented to demonstrate the performance of the estimation procedure and the performance of the new statistic compared to the LR and W tests. The last section provides the concluding remarks.

Cognitive Diagnosis Modeling

Cognitive Diagnosis Modelings are multidimensional, categorical latent-trait models developed primarily for identifying which attributes (e.g., skills, mental disorders, competencies)

DOI: 10.1027/1614-2241/a000131

are mastered and which ones do not (see, e.g., Rupp & Templin, 2008, for an overview of these models). For an assessment diagnosing K attributes, examinees are grouped into 2^K latent classes. Latent classes are represented by an attribute vector denoted by $\boldsymbol{\alpha}_l = (\alpha_{l1}, \alpha_{l2}, \ldots, \alpha_{lK})$, where $l = 1, \ldots, 2^K$. Specifically, $\alpha_{lK} = 1$ or 0 represents mastery or nonmastery of attribute k, respectively. In each latent class, examinees all have the same probability of success on a particular item j, denoted by $P(X_j = 1|\boldsymbol{\alpha}_l) = P_j(\boldsymbol{\alpha}_l)$. In other contexts (e.g., measurement of psychological disorders), $P_j(\boldsymbol{\alpha}_l)$ indicates the probability of item endorsement. The attributes that are required to correctly answer each item are defined in a $J \times K$ matrix, commonly known as Q-matrix (Tatsuoka, 1990), where J is the test length.

Several general models that encompass reduced CDMs have been proposed, including the above-mentioned G-DINA model. The G-DINA model is a generalization of the *deterministic inputs, noisy,* "*and*" *gate* (DINA; Haertel, 1989) model that describes the probability of success on item j in terms of the sum of the effects of the attributes involved and their corresponding interactions. Let the number of required items for item j be denoted by K_j^*. In this model, latent classes are sorted into $2^{K_j^*}$ latent groups. Each of these latent groups represents one reduced attribute vector $\boldsymbol{\alpha}_{lj}^*$. The probability of success associated to $\boldsymbol{\alpha}_{lj}^*$ is defined as

$$P(\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}, \quad (1)$$

where δ_{j0} is the intercept or baseline probability for item j, δ_{jk} is the main effect due to α_k, $\delta_{jkk'}$ is the interaction effect due to α_k and $\alpha_{k'}$, and $\delta_{j12\ldots K_j^*}$ is the interaction effect due to $\alpha_1,\ldots,\alpha_{K_j^*}$. Thus, there are $2^{K_j^*}$ parameters to be estimated for item j.

By constraining the parameters of the saturated model, de la Torre (2011) has shown that some of the commonly used reduced CDMs can be obtained, including the DINA model and the *additive* CDM (*A*-CDM; de la Torre, 2011). To compare the different models in a more straightforward manner, this article uses $\boldsymbol{\varphi}_j$ to represent reduced model item parameters across all reduced CDMs. Namely, φ_{j0} is the intercept for item j, φ_{jk} is the main effect due to α_k, and $\varphi_{j12\cdots K_j^*}$ is the interaction effect due to $\alpha_1, \ldots, \alpha_{K_j^*}$. The DINA model is a conjunctive model, that is, an examinee needs to have mastered all required attributes to correctly answer a particular item. As such, the DINA model separates examinees into two latent groups for each item: one group with examinees who have mastered all attributes required by the item and one group with examinees lacking at least one. The probability of correct response is represented by the DINA model as follows:

$$P(\boldsymbol{\alpha}_{lj}^*) = \varphi_{j0} + \varphi_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ljk}. \quad (2)$$

Therefore, the DINA model has two parameters per item and is deduced from the G-DINA model by setting to zero all terms except for δ_{j0} and $\delta_{j12\ldots K_j^*}$ to zero. For the *A*-CDM, all the interaction terms are dropped. The item response function is given by

$$P(\boldsymbol{\alpha}_{lj}^*) = \varphi_{j0} + \sum_{k=1}^{K_j^*} \varphi_{jk}\alpha_{jlk}. \quad (3)$$

This is the G-DINA without the interaction terms, and it shows that mastering attribute α_{lk} raises the probability of success on item j by φ_{jk}. There are $K_j^* + 1$ parameters for item j in the *A*-CDM. In this respect, the DINA model involves a conjunctive process, whereas the *A*-CDM involves an additive process. Figure 1 gives a graphical representation of an item requiring two attributes when it conforms to the DINA, *A*-CDM, or the G-DINA model. As can be observed from Figure 1, in the DINA model latent classes are sorted into two latent groups. Examinees who have mastered all attributes required by the item have a probability of correct response equal to $\varphi_{j0} + \varphi_{j12\cdots K_j^*}$. Examinees lacking at least one attribute will have a probability of correct response equal to the baseline probability (i.e., φ_{j0}). In the case of the *A*-CDM, each attribute has a main impact. For example, examinees mastering only the first attribute will have a probability of success equal to $\varphi_{j0} + \varphi_{j1}$.

Model Comparison in CDM

Each CDM assumes a different cognitive process involved in responding to an item (e.g., conjunctive or additive). The task in model selection is to select the model that is the best fit to the data. For nested CDMs, model selection at the item level can be done using the three common tests for assessing relative fit (Buse, 1982): likelihood ratio (LR), Wald (W), and Lagrange multiplier (LM) tests. In all the three cases, the statistic is assumed to be asymptotically χ^2 distributed with $2^{K_j^*} - p$ degrees of freedom, where p is the number of parameters of the reduced model.

To investigate the finite sample performance of these tests, Sorrel et al. (2017) conducted a simulation study. Overall the Type I error and power comparisons favored LR and W tests over the LM test. LR was found to be relatively more robust than the W test. However, the appealing advantage of using the W test is that it required only the

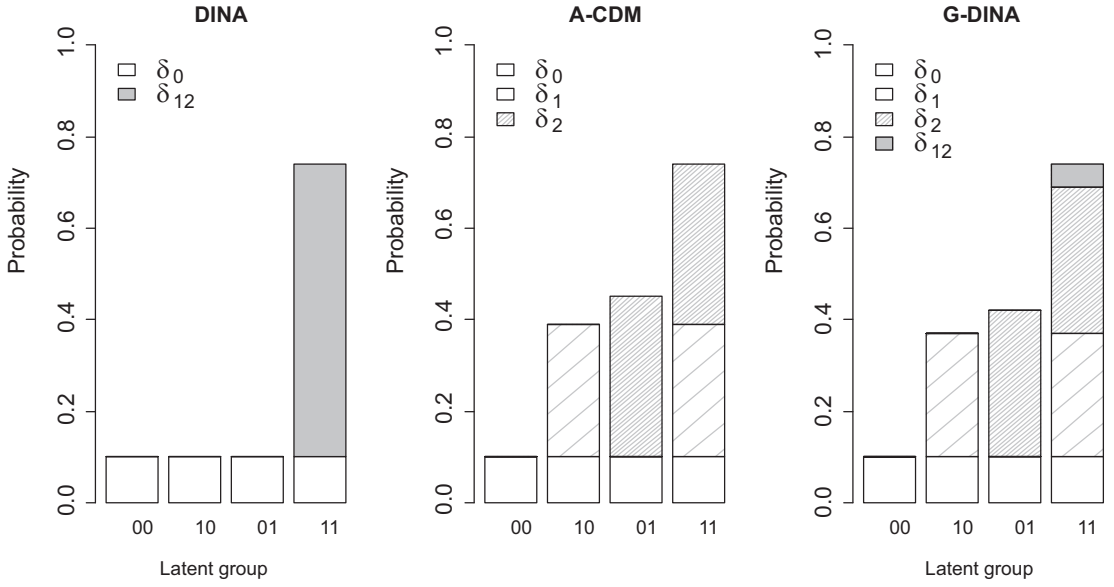

Figure 1. This figure depicts the probability of correctly answering an item requiring two attributes for the DINA, *A*-CDM, and G-DINA models. Item parameters are denoted by δ.

unrestricted model (i.e., G-DINA) to be estimated. In contrast, the LR test required $J^* \cdot NR + 1$ models to be estimated, where J^* is the number of items measuring more than one attribute and NR is the number of reduced models to be tested. In this study, we propose an approximation to the LR test, 2LR, which also has the appealing advantage of only requiring the G-DINA model to be estimated. In the following, we will describe how the 2LR test is computed.

Approximation to the LR test

The LR test is a statistical test used to compare the goodness-of-fit of two models, one of which is nested in the other. Because adding additional parameters to a more general model will always result in a higher likelihood, CDMs with general formulations will provide a better fit to the data. The LR test provides one objective criterion for evaluating if the more general model fits a particular dataset significantly better. In the traditional implementation of the LR test in the CDM context, the more general model, the G-DINA model, is estimated for all the items. This model is compared with a reduced model fitted to a target item, whereas the G-DINA model is fitted to the rest of the items. Both model specifications are estimated and the LR statistic is computed as twice the difference in the log-likelihoods. The application of the LR test requires comparing the different combinations of the models. To obtain the likelihood of a model, both item parameter and posterior distribution estimates are needed. Rojas, de la Torre, and Olea (2012) found that the attribute classification accuracy of the G-DINA model is the best when the underlying model is not known. de la Torre and Chen (2011) introduced a procedure for estimating the reduced model item parameters using the attribute classification obtained with the G-DINA. Let us review their proposal.

Two-Step Estimation Procedure

de la Torre and Chen (2011) originally introduced an alternative estimation procedure that uses the G-DINA estimates for efficiently estimating the parameters of several reduced CDMs. This method is referred to as two-step estimation procedure because the estimation of the item parameters (i.e., $\boldsymbol{\varphi}_j$) for the reduced CDMs is done in two steps. The first step involves estimating the G-DINA model parameters, $P_j = \left\{\boldsymbol{P}\left(\boldsymbol{\alpha}_{lj}^*\right)\right\}$. The second step involves computing the corresponding $\boldsymbol{\varphi}_j$ of the reduced models. de la Torre (2011) showed that $\boldsymbol{P}_j$ can be estimated using an expectation-maximization (EM) implementation of the marginal maximum likelihood (MML) estimation. Briefly, it can be shown that the MML estimate of the parameter $P\left(\boldsymbol{\alpha}_{lj}^*\right)$ is given by

$$\hat{P}\left(\boldsymbol{\alpha}_{lj}^*\right) = \frac{R_{\alpha_{lj}^*}}{I_{\alpha_{lj}^*}}, \tag{4}$$

where $I_{\alpha_{lj}^*}$ and $R_{\alpha_{lj}^*}$ are the expected number of examinees and correct responses in the latent group $\boldsymbol{\alpha}_{lj}^*$, respectively.

Once $\mathbf{P}_j$ has been estimated, item parameters φ_j can be obtained through some linear transformations or maximization processes. For DINA model, a $2^{K_j^*} \times p$ design matrix $\mathbf{M}$

can be used to linearly transform the G-DINA model parameters into reduced model item parameters, where p is the number of model parameters. To illustrate, let $K_j^* = 2$. The saturated design matrix is

$$\mathbf{M}_{4\times 4}^{(\mathrm{S})} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}. \quad (5)$$

In deriving the reduced models, subsets of functions of subsets of the columns of $\mathbf{M}^{(\mathrm{S})}$ are used. For example, the design matrix for the DINA model would be

$$\mathbf{M}_{4\times 2} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}. \quad (6)$$

The design matrix for the DINA model indicates that all elements of $\mathbf{P}_j$ contain $\varphi_{j\mathrm{o}}$ whereas only the last element contains $\varphi_{j12\cdots K_j^*}$. Several elements of $\mathbf{P}_j$ need to be combined to obtain $\boldsymbol{\varphi}_j$. These elements are differentially weighted to account for the relative size of the latent classes. DINA model estimates are obtained by

$$\hat{\boldsymbol{\varphi}}_j = (\mathbf{M}'\mathbf{W}\mathbf{M})^{-1}\mathbf{M}'\mathbf{W}\hat{\mathbf{P}}_j, \quad (7)$$

where $\mathbf{W}$ is a diagonal matrix $\mathbf{W}_{2^{K_j^*}\times 2^{K_j^*}} = \{I_{\boldsymbol{\alpha}_{lj}^*}\}$ and $\hat{\mathbf{P}}_j = \{\hat{P}(\boldsymbol{\alpha}_{lj}^*)\}$.

For *A*-CDM, however, the design matrix cannot be used because $\boldsymbol{\varphi}_j$ cannot be expressed as a simple linear combination of the elements of $\mathbf{P}_j$. Instead, the parameter estimates can be obtained by maximizing the likelihood of $\boldsymbol{\varphi}_j$ given $\mathbf{R}_j = \{\mathrm{R}_{\boldsymbol{\alpha}_{lj}^*}\}$ and $\mathbf{I}_j = \{\mathrm{I}_{\boldsymbol{\alpha}_{lj}^*}\}$ obtained in the first step as follows:

$$\mathbf{L}(\boldsymbol{\varphi}_j|\mathbf{R}_j, \mathbf{I}_j) = \prod_{l=1}^{2^{K_j^*}} P^{(R)}(\boldsymbol{\alpha}_{lj}^*)^{R_{\boldsymbol{\alpha}_{lj}^*}} \left[1 - P^{(R)}(\boldsymbol{\alpha}_{lj}^*)\right]^{\left(I_{\boldsymbol{\alpha}_{lj}^*} - R_{\boldsymbol{\alpha}_{lj}^*}\right)}, \quad (8)$$

where $P^{(R)}(\boldsymbol{\alpha}_{lj}^*)$ is the probability of success implied by the reduced model. In this article we explore how this estimation procedure can be used as a basis in efficiently computing an approximation to the LR test.

Two-Step Likelihood Ratio Test

Item-level maximum likelihoods for the saturated and reduced models can be computed based on the estimated item parameters and attribute distribution. Item parameters are those estimated for G-DINA and the reduced model (i.e., DINA or *A*-CDM), whereas the attribute distribution is obtained in the first step based on the G-DINA model. Comparing the two marginalized likelihoods using a LR test can be useful to find out if a reduced model is appropriate for those items measuring more than one attribute. We proposed the 2LR test as an efficient way of computing the statistic, and it is computed as

$$\begin{aligned} 2\mathrm{LR}_j &= 2\left[\log L(\mathbf{P}_j|\mathbf{R}_j, \mathbf{I}_j) - \log L\left(\boldsymbol{\varphi}_j|\mathbf{R}_j, \mathbf{I}_j\right)\right] \\ &\sim \chi^2\left(2^{K_j^*} - p\right), \end{aligned} \quad (9)$$

where $\mathbf{P}_j$ is the vector of GDINA item parameters and $\boldsymbol{\varphi}_j$ is the vector of reduced model parameters for item j. The likelihood function that is employed is the one represented in (8). Compared to the LR test, only one model (i.e., G-DINA) is estimated. Given that the two-step estimation procedure is the basis of the new statistic, it is pivotal to ensure its accuracy under plausible scenarios.

Method

A simulation study was conducted to assess the accuracy of the two-step item parameter estimates and performance of the 2LR test compared to the LR and W tests. Four factors were varied and their levels were chosen to represent realistic scenarios. These factors were: (1) generating model (*MOD*; DINA model and *A*-CDM); (2) test length (*J*; 30 and 60 items); (3) sample size (*N*; 500, 1,000, and 2,000 examinees); and (4) item quality or discrimination, defined as the difference between the maximum and the minimum probabilities of correct response according to the attribute latent profile (IQ; .40, .60, and .80).

The probabilities of success for individuals who mastered none of the required attributes were fixed to .30, .20, and .10 for the low, medium, and high item quality conditions, respectively; the corresponding probabilities for those who mastered all of the required attributes were fixed to .70, .80, and .90. For the *A*-CDM, an increment of $.40/K_j^*$, $.60/K_j^*$, and $.80/K_j^*$ was associated with each attribute mastery for the low, medium, and high item quality conditions, respectively. The number of attributes was fixed to $K = 5$. The Q-matrix used in simulating the response data and fitting the models is given in Table 1. This Q-matrix was constructed such that each attribute appears alone, in a pair, or in a triple the same number of times as other attributes. For $J = 60$, each item was used twice.

For each of the 36 factor combinations, 200 datasets were generated and DINA, *A*-CDM, and G-DINA models were fitted. We evaluated whether the two-step algorithm is comparable, in terms of estimation accuracy or variability, to the standard EM-MML algorithm. For comparison of estimation accuracy, we computed the bias, $\hat{\varphi} - \varphi$; for

Table 1. Simulation study Q-matrix for the $J = 30$ conditions

	Attribute				
Item	α_1	α_2	α_3	α_4	α_5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
6	1	0	0	0	0
7	0	1	0	0	0
8	0	0	1	0	0
9	0	0	0	1	0
10	0	0	0	0	1
11	1	1	0	0	0
12	1	0	1	0	0
13	1	0	0	1	0
14	1	0	0	0	1
15	0	1	1	0	0
16	0	1	0	1	0
17	0	1	0	0	1
18	0	0	1	1	0
19	0	0	1	0	1
20	0	0	0	1	1
21	1	1	1	0	0
22	1	1	0	1	0
23	1	1	0	0	1
24	1	0	1	1	0
25	1	0	1	0	1
26	1	0	0	1	1
27	0	1	1	1	0
28	0	1	1	0	1
29	0	1	0	1	1
30	0	0	1	1	1

Note. The Q-matrix for the $J = 60$ conditions is doubled from this Q-matrix.

comparison of estimation variability, empirical *SE*s (i.e., standard deviations across replications) were computed.

The W, LR, and 2LR tests were computed for each dataset. In addition to assessing whether the 2LR test is a good approximation to the LR test, we also compared the performance of the 2LR and W tests in terms of Type I error and power. Type I error was computed as the proportion of times H_0 was rejected when the fitted reduced model is true; power was computed as the proportion of times that a wrong reduced model was rejected. The significance level was set at .05. With 200 replicates, the 95% confidence interval for the Type I error is given by $.05 \pm 1.96\sqrt{.05(1-.05)/200} = [.02, .08]$. A conservative performance (i.e., Type I error < .02) might be a good characteristic provided the power is not affected. For the purposes of this work, a statistical procedure was considered to be "good" if it had a Type I error within the [0, .08] interval and a relatively high power (> .80). The code used in this article was written in R. Some functions included in the GDINA (Ma & de la Torre, 2016) package were employed. The R code can be requested by contacting the corresponding author.

Results

Two-Step Estimation Procedure

Due to space limits, only the results of comparison in the worst ($N = 500$, $J = 30$, and IQ = LD) and best conditions ($N = 2{,}000$, $J = 60$, and IQ = HD) are presented in Tables 2 and 3. Items 1, 11, and 21, and 1, 21, and 41 are selected for $J = 30$ and $J = 60$, respectively. These are the same items that represent $K_j^* = 1$, 2, and 3 for both Q-matrices. In the case of the DINA model, we study the recovery of the probability of correct response in the two possible latent groups (i.e., φ_{jo} and $\varphi_{jo} + \varphi_{j12\ldots K_j^*}$). In the case of the *A*-CDM, we study the recovery of the baseline probability and the probability of correct response for examinees mastering only the first attribute (i.e., φ_{jo} and $\varphi_{jo} + \varphi_{j1}$).

In the worst condition, differences in terms of bias and empirical *SE* between the two algorithms were small, ranging from −.015 to .046, .010 being the mean and 0.012 the standard deviation. Not surprisingly, there was almost no difference between the two algorithms in the best condition – the largest absolute difference was 0.001. Considering both conditions, we can safely conclude that the differences of estimation accuracy and variability between the EM-MMLE and two-step algorithms were negligible. It should be noted that empirical SEs associated to the *A*-CDM estimates were usually larger compared to the DINA estimates. For example, this can be observed for the two-step estimates for item 21 in the worst condition. Empirical *SE*s for the *A*-CDM probabilities were .067 and .101. In the same condition, empirical *SE*s for the DINA probabilities were .038 and .045.

Two-Step Likelihood Ratio Test

Descriptive Analysis

All the item fit statistics were highly correlated. The Pearson correlation coefficients ranged from .97 to .99. Average computing time was recorded separately for each statistic. As an example, we found that in one of the most extreme conditions (i.e., $N = 2{,}000$, $J = 60$, and IQ = LD) the LR and 2LR tests took 475.03 and 1.61 seconds per replicate, respectively. In other words, the 2LR test was 295 times faster than the LR test.

Table 2. Selected item estimates for the DINA model

					Bias		Empirical standard error	
N	*J*	IQ	Item	Estimation algorithm	φ_{j0}	$\varphi_{j0}+\varphi_{j12\cdots K_j^*}$	φ_{j0}	$\varphi_{j0}+\varphi_{j12\cdots K_j^*}$
2000	60	HD	1	EM-MMLE	.000	.001	.012	.008
				Two-step	.000	.001	.012	.008
			21	EM-MMLE	.001	.000	.009	.009
				Two-step	.001	.000	.009	.009
			41	EM-MMLE	−.001	.000	.009	.010
				Two-step	−.001	.000	.009	.010
500	30	LD	1	EM-MMLE	−.010	−.001	.073	.030
				Two-step	.036	.001	.081	.034
			11	EM-MMLE	.003	.001	.043	.031
				Two-step	.029	.008	.052	.039
			21	EM-MMLE	.001	−.002	.030	.039
				Two-step	.018	.003	.038	.045

Notes. Generating values for the probabilities in the low discrimination (high discrimination) conditions were .30 (.10) and .70 (.90) for φ_{j0} and $\varphi_{j0}+\varphi_{j12\cdots K_j^*}$, respectively. *N* = sample size; *J* = test length; IQ = item quality; HD = high discrimination; LD = low discrimination.

Table 3. Selected item estimates for the A-CDM

					Bias		Empirical standard error	
N	*J*	IQ	Item	Estimation algorithm	φ_{j0}	$\varphi_{j0}+\varphi_{j1}$	φ_{j0}	$\varphi_{j0}+\varphi_{j1}$
2000	60	HD	1	EM-MMLE	−.001	.001	.012	.007
				2-step	−.001	.001	.012	.007
			21	EM-MMLE	−.001	.000	.016	.022
				2-step	−.001	−.001	.016	.021
			41	EM-MMLE	.000	−.001	.017	.025
				2-step	.000	−.002	.017	.025
500	30	LD	1	EM-MMLE	.003	−.004	.082	.040
				2-step	.033	−.005	.086	.040
			11	EM-MMLE	.001	.002	.079	.097
				2-step	.022	.022	.075	.098
			21	EM-MMLE	−.008	−.010	.060	.116
				2-step	.004	.003	.067	.101

Notes. Generating values for the probabilities in the low discrimination (high discrimination) conditions were .30 (.10) for φ_{j0} and .70, .50, and .43 (.90, .50, and .37) for $\varphi_{j0}+\varphi_{j1}$ for items 1, 11, 21, respectively. *N* = sample size; *J* = test length; IQ = item quality; HD = high discrimination; LD = low discrimination.

Type I Error

Type I error study results are presented in Table 4. The LR and 2LR tests were generally preferable to the W test. Type I error for the 2LR test was very similar to that obtained for the LR test. With the exception of low discriminating items, the 2LR and LR tests had an acceptable Type I error. LR and 2LR Type I error was close to the nominal level with low quality items when the sample size and test length were large (N = 2,000 and J = 60). 2LR Type I error was particularly good for DINA generated data. It was the only one lower than the upper limit of the confidence interval with medium quality items and a small sample size and test length (N = 500 and J = 30). The W test generally required a larger sample size. For example, W Type I error rate was inflated with medium quality items and small sample size (N = 500 and 1,000).

Power

Power study results are presented in Table 5. Power results should always be interpreted with some caution because power comparisons require equal Type I error. More liberal tests have a higher power because they tend to overestimate the significance. Power for all statistics was always higher than 0.80 and close to 1.00 in the high and medium discrimination conditions. In the case of the low quality items conditions, a large number of examinees (i.e., 1,000 or 2,000) or items (i.e., 60) were needed to reach acceptable values (i.e., > 0.80). 2LR power tended to be

Table 4. Type I error of the item fit statistics (LR, 2LR, and W) for the DINA and A-CDM models

Factors			DINA			*A*-CDM		
IQ	*J*	*N*	LR	2LR	W	LR	2LR	W
HD	30	500	.066	.022	.053	.058	.029	.079
		1,000	.062	.022	.090	.054	.029	.063
		2,000	.058	.022	.069	.048	.027	.054
	60	500	.061	.017	.055	.052	.016	.061
		1,000	.060	.017	.076	.051	.016	.055
		2,000	.051	.015	.062	.047	.015	.049
MD	30	500	.101	.075	.163	.145	.110	.233
		1,000	.068	.065	.109	.074	.083	.116
		2,000	.062	.060	.078	.053	.079	.067
	60	500	.070	.026	.105	.069	.033	.098
		1,000	.061	.026	.079	.059	.034	.070
		2,000	.050	.020	.060	.054	.031	.057
LD	30	500	.358	.443	.595	.374	.235	.581
		1,000	.223	.334	.430	.297	.290	.519
		2,000	.131	.278	.235	.224	.316	.371
	60	500	.199	.131	.323	.302	.133	.418
		1,000	.101	.096	.190	.156	.144	.262
		2,000	.071	.082	.102	.075	.118	.116

Notes. Shaded cells correspond to values in the [.00, .08] interval. IQ = item quality; *J* = test length; *N* = sample size; LR = likelihood ratio test; 2LR = two-step likelihood ratio test; W = Wald test; HD = high discrimination; MD = medium discrimination; LD = low discrimination.

Table 5. Power of the item fit statistics (LR, 2LR, and W) for the DINA and A-CDM models

Factors			Generating, true model: DINA Fitted, false model: *A*-CDM			Generating, true model: *A*-CDM Fitted, false model: DINA		
IQ	*J*	*N*	LR	2LR	W	LR	2LR	W
HD	30	500	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		1,000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	1.000
		2,000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	60	500	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		1,000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		2,000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
MD	30	500	1.000	1.000	1.000	.860	**.956**	.938
		1,000	**1.000**	1.000	1.000	**.994**	**.999**	.996
		2,000	**1.000**	**1.000**	1.**000**	**1.000**	**1.000**	**1.000**
	60	500	**1.000**	**1.000**	1.000	**.971**	**.979**	.980
		1,000	**1.000**	**1.000**	1.**000**	**1.000**	**1.000**	**1.000**
		2,000	**1.000**	**1.000**	1.**000**	**1.000**	**1.000**	**1.000**
LD	30	500	.595	.748	.759	.526	.776	.799
		1,000	.819	.952	.905	.589	.893	.837
		2,000	.979	.999	.987	.721	.959	.892
	60	500	.835	.916	.914	.533	.706	.749
		1,000	.984	.996	.991	.722	.906	.849
		2,000	**1.000**	1.000	1.000	**.963**	.995	.975

Notes. Shaded cells correspond to values in the [.80, 1.00] interval. Values shown in bold correspond to conditions where the actual Type I error was within the [.00, .08] interval. IQ = item quality; *J* = test length; *N* = sample size; LR = likelihood ratio test; 2LR = two-step likelihood ratio test; W = Wald test; HD = high discrimination; MD = medium discrimination. LD = Low discrimination.

higher that of the LR and W tests. For example, this was usually the case in the medium item quality conditions. It should be noted that in these conditions the 2LR Type I error was within the [0, .08] interval. In addition, it is important to note that, in the case of *A*-CDM generated data, 2LR power was much higher than that of the LR test in the low quality conditions, being .68 and .87 the marginal means associated to the LR and 2LR tests, respectively.

Discussion

Model-fit has received greater attention in the recent CDM literature (e.g., Chen et al., 2013; de la Torre & Lee, 2013; Hansen, Cai, Monroe, & Li, 2016; Liu et al., 2016; Sorrel et al., 2017). This an important area of research because proper application of a statistical model requires the assessment of model-data fit. Different reduced CDMs with different assumptions have been proposed in the literature. For example, the DINA model assumes a conjunctive process in that only individuals who master all required attributes are expected to correctly answer the item and the *A*-CDM assumes that the different attributes measured by the item contribute independently to the probability of correctly answer the item. A critical concern is in selecting the most appropriate model for each item from the available CDMs. To do so, several tests for model comparison have been employed, which include LR and the W tests. Although it has been found in the CDM context that the LR test is relatively more robust than the W test (Sorrel et al., 2017), the current implementation of the LR test is very time consuming, given that it requires to calibrate many different models and compare them to the general model. For this reason, the W test is generally preferred (e.g., de la Torre et al., 2015; Ravand, 2016) and is the one implemented in the software available (e.g., the CDM and GDINA packages in R; Ma & de la Torre, 2016; Robitzsch, Kiefer, George, & Uenlue, 2016).

In this work, we introduce an efficient approximation to the LR test, 2LR, based on a two-step estimation procedure under the G-DINA model framework originally introduced by de la Torre and Chen (2011). Results indicate that this two-step estimation procedure is comparable in terms of estimation accuracy and variability to the standard procedure based on EM-MMLE. Mean absolute differences and empirical standard errors produced by the two algorithms were very similar, even in the worst conditions. This shows that the estimates based on the two-step estimation procedure can be used to develop the approximation to the LR test.

The simulation study results allow us to draw several conclusions about the performance of the LR, 2LR, and W tests. First, the LR and 2LR tests were highly correlated. The performance of the 2LR test was very similar to that of the LR test. However, the computation of the 2LR test was remarkably faster. Secondly, the LR and 2LR tests were found to perform better than the W test. Thirdly, there was a large effect of the item quality. Type I error was close to the nominal level when the item quality was medium or high. In the poor discriminating item conditions, Type I error was inflated but in the case of the LR and 2LR tests this could be compensated by increasing the number of items or the sample size. Power decreased in the poor discriminating conditions. It is noteworthy that 2LR power was the least affected in these conditions and tended to be high. In sum, the 2LR test can be recommended for use in empirical research.

Following are some of the limitations of the current study and several avenues for future research. First, although not considered here, there is an additional reason to prefer the LR test over the W test – the LR test does not require the standard errors of the item parameter estimates, whereas the W test does. Future studies should explore the advantages of this feature. Second, all items were simulated to have the same discrimination power. This might not be feasible in practice. Finally, we focus on the DINA and *A*-CDM models and five attributes. Future studies might manipulate the number of attributes and try to extend this results to other models such as the *deterministic inputs, noisy "or" gate* (DINO) model (Templin & Henson, 2006), the linear logistic model (LLM; Maris, 1999), and the *reduced reparameterized unified model* (R-RUM; Hartz, 2002).

Acknowledgments

This research was supported by Grant PSI2013-44300-P (Ministerio de Economia y Competitividad and European Social Fund).

References

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange Multiplier tests: An expository note. *American Statistician, 36*, 153–157. doi: 10.2307/2683166

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*, 123–140. doi: 10.1111/j.1745-3984.2012.00185.x

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199. doi: 10.1007/s11336-011-9207-7

de la Torre, J., & Chen, J. (2011, April). *Estimating different reduced cognitive diagnosis models using a general framework*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*, 355–373. doi: 10.1111/jedm.12022

de la Torre, J., van der Ark, L. A., & Rossi, G. (2015). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*. Advance Online Publication. doi: 10.1177/0748175615569110

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333–352. doi: 10.1111/j.1745-3984.1989.tb00336.x

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *British Journal of Mathematical and Statistical Psychology, 69*, 225–252. doi: 10.1111/bmsp.12074

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing, 11*, 144–177. doi: 10.1080/15305058.2010.534571

Liu, Y., Tian, W., & Xin, T. (2016). An application of M2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics, 41*, 3–26. doi: 10.3102/1076998615621293

Ma, W., & de la Torre, J. (2016). *GDINA: The generalized DINA model framework*. R package version 0.13.0. Retrieved from http://CRAN.R-project.org/package=GDINA

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187–212. doi: 10.1007/BF02294535

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment, 34*, 782–799. doi: 10.1177/0734282915623053

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2016). *CDM: Cognitive Diagnosis Modeling*. R package version 4. 8-0. Retrieved from http://CRAN.R-project.org/package=CDM

Rojas, G., de la Torre, J., & Olea, J. (2012, April). *Choosing between general and specific cognitive diagnosis models when the sample size is small*. Paper presented at the meeting of the National Council on Measurement in Education, Vancouver, Canada.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*, 219–262. doi: 10.1080/15366360802490866

Sorrel, M. A., Abad, F. J., Olea, J., Barrada, J. R., & de la Torre, J. (2017). Inferential item fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement*. Advance online publication. doi: 10.1177/0146621617707510

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods, 19*, 506–532. doi: 10.1177/1094428116630065

Tatsuoka, K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Safto (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305. doi: 10.1037/1082-989X.11.3.287

Published online June 1, 2017

Miguel A. Sorrel
Department of Social Psychology and Methodology
Universidad Autónoma de Madrid
Ciudad Universitaria de Cantoblanco
Madrid 28049
Spain
miguel.sorrel@uam.es